mvBIMBAM

June 17, 2014

1 Preamble

The `mvBIMBAM`, a version of `BIMBAM` (`http://stephenslab.uchicago.edu/software.html`) [1] for multivariate association analysis, implements a Bayesian approach for genetic association analysis of multiple related phenotypes. Note also that although this version of BIMBAM can also do imputation etc, we ask that you instead use the official released version for all this kind of work; this version is really for testing the multivariate phenotype option only. These instructions are just for how to run the multivariate analyses: you will need to consult the regular `BIMBAM` instructions (`http://www.haplotype.org/download/bimbam-manual.pdf`) for input file format etc.

2 Brief Statistical Background

A multivariate phenotype analysis can be performed using the `-mph` option, as described below. Note that this option currently only allows single-SNP analyses (i.e. the SNPs in a genome-wide scan will be analyzed one at a time).

The analysis is based on the multivariate regression model:

$$Y = \mu + GB + E \tag{1}$$

where $Y(n \times d)$ is a matrix of d phenotypes on each of n individuals; $\mu(n \times 1)$ is an intercept vector; $G(n \times 1)$ is a vector of genotypes (usually coded 0, 1 and 2) measured on the same individuals; $B(1 \times d)$ is a vector of regression coefficients relating the phenotypes to the genotypes; and $E(n \times d)$ is a matrix of error terms, whose rows we assume to be independent and identically distributed as $N_d(0, V)$ for some unknown covariance matrix $V(d \times d)$.

Please read [2] for details of the method. In brief, the approach attempts to partition the response variables Y into three groups – denoted U, D and I – consisting of those variables that are Unaffected by G, those that are Directly affected by G, and those that are Indirectly affected by G through the variables in D. Here the terms "direct" and "indirect" are meant statistically and not molecularly. For example, we do not mean a "direct" effect of G on Y_D to be interpreted as necessarily implying a direct molecular mechanism linking the two (although such a mechanism could be one possible explanation).

More formally the idea is to consider the set of models H_γ, where the index $\gamma = (U, D, I)$ runs through partitions of the coordinates $\{1, \dots, d\}$. That is, U, D, and I are disjoint subsets of $\{1, \dots, d\}$, whose union is $\{1, \dots, d\}$. Under model H_γ we assume that Y_U is independent of X, and Y_I is conditionally independent of X given Y_D. This gives

$$p_\gamma(Y|X) = p_\gamma(Y_U)p_\gamma(Y_D|Y_U, X)p_\gamma(Y_I|Y_U, Y_D). \quad (2)$$

BIMBAM computes the Bayes Factor which measures the support for partition γ compared wtih the global null that all the phenotypes are unassociated with G (i.e. all the phenotypes are in U). It then summarizes both the overall evidence against the null, as well as the posterior probability that each coordinate of Y is associated with G (see discussion of output files below).

3 Performing an analysis

Because the method is based on an assumption of multivariate normality (within each genotype class), we recommend you consider i) quantile normalizing each phenotype to the quantiles of a standard normal $N(0, 1)$ distribution to provide some robustness to the normality assumption; ii) plot scatterplots of every pair of phenotypes to check the multivariate normal assumption doesn't look too bad.

The software can be run two different ways, using either the `-mph 1` or `-mph 2` options.

3.1 A fast initial analysis

The first way, using `-mph 1` is fast but less interesting: in effect it just does a simple multivariate test of the null hypothesis vs a general multivariate alternative. It can handle a moderately large number of phenotypes and be run on a whole-genome scale (we haven't tested the limits, but in principle this analysis should not be much more intensive than univariate analyses). It is intended primarily to be used, if necessary, as an initial filtering step, before running the more computationally-intensive `-mph 2` analyses on those SNPs with the strongest signals. If you have only a handful of phenotypes then you may be able to run the `-mph 2` analysis genome wide on all SNPs, and have no need for this first type of analysis.

Here is an example of how to run this analysis on the example files

```
./bimbam -g example/test.geno.txt -p example/test.multi-pheno.txt
-o test.mph1 -f 3 -mph 1 -A 0.1 -A 0.2
```

The number of phenotypes is specified using the `-f` option. In this example, the phenotype input file (`test.multi-pheno.txt`) contains three columns each of which corresponds to each phenotype (note: currently BIMBAM does not allow missing phenotypes for the multivariate phenotype analysis). Like the univariate phenotype analysis, the `-A` option is used to specify multiple values for σ_a (see instructions for regular BIMBAM for details). The values 0.1 and 0.2 are not too bad as defaults, but you might want to check for robustness to a wider range of values (See also [3] for further discussion).

3.1.1 The output file

This command generates one output file (`test.mph1.mph.BFs.txt`).

```
<test.mph1.mph.BFs.txt>

rs1 NA -0.00135 +0.18436 -0.11280 -0.08671
rs13 NA +0.00210 +0.21745 -0.11803 -0.09825
```

The first column is the SNP name. The second column is missing (NA), as it is filled in only in the second kind of analysis (see below). The third column contains the $\log_{10}$ of the Bayes Factor for testing for association between the multivariate phenotype and genotype. (Formally it is the BF for the partition γ in which all variables are "directly" affected by G in the model described above; it is by only considering this one partition that this analysis is so fast). Large values of this BF indicate SNPs that have the strongest evidence for association with the multivariate phenotype. Just as in a univariate analysis, this BF can be weighed against the Prior Odds to obtain the Posterior Odds for association. See [3] for more discussion. As a rough guide, in a genetic association study, a $\log_{10}$ value of ¿ 4 might be considered moderate evidence for association, and $\log_{10}$ value of ¿ 6 might be considered strong evidence.

Remaining columns give the $\log_{10}$ univariate BFs for each of the phenotypes considered in turn.

3.2 A more detailed multivariate analysis

The `-mph 1` analysis is fast, because it does not consider all the different possible partitions of phenotypes into the different categories U, D and I. As a result it does not allow you to ask interesting questions such as *which* of the phenotypes is associated with each SNP. To do this more detailed analysis use the `-mph 2` option, using for example

```
./bimbam -g example/test.geno.txt -p example/test.multi-pheno.txt
-o test.mph2 -f 3 -mph 2 -A 0.1 -A 0.2
```

3.2.1 The output files

The `-mph 2` analysis generates three output files.

`<test.mph2.mph.BFs.txt>`

```
rs1 +0.02371 -0.00135 +0.18436 -0.11280 -0.08671
rs13 +0.03043 +0.00210 +0.21745 -0.11803 -0.09825
```

This file, which should be the first one you check to identify *which* SNPs have the strongest evidence for assocation, is just the same as for `-mph 1`, except that the second column is filled in. This second column contains a $\log_{10}$ Bayes Factor that tests for *any* association between SNP and the phenotypes, averaging over all possible partitions (γ) of the phenotypes into the different groups. Once again this BF should be weighed against the small prior odds of association, so you generally need a big value to claim strong evidence for association. In general this BF will be highly correlated with simpler BF computed in the third column; however it should be more effective at identifying cases where a SNP is associated with only a small number of the phenotypes (i.e. the BF in the second column will be larger than the BF in the third column for those SNPs that are associated with only a few of the phenotypes).

`<test.mph2.mph.prob.txt>`

```
rs1  +0.24934 +0.63068 +0.38824 +0.42576 +0.38073 +0.43003
rs13 +0.23739 +0.64702 +0.39210 +0.41642 +0.38571 +0.42168
```

This second output file allows you to identify, in more detail *which* phenotypes are likely associated with each SNP, conditional on an overall association with at least one phenotype. Thus, you should only bother with this file for those SNPs that you believe are likely associated with the phenotypes (usually, and ideally, because they have a large BF in the first file, but in some cases possibly also because of external data connecting that SNP with these phenotypes).

Specifically the file allows you to compute the marginal posterior probabilities for each phenotype being in each of the three groups, unaffected, directly affected and indirectly affected for each SNP (conditional on at least one phenotype being associated with the SNP). The first column contains the SNP name. Then each phenotype is represented by two adjacent columns containing the marginal posterior probabilities for the phenotype being unaffected (the former) and directly affected (the latter). The marginal posterior probability for the phenotype being indirectly affected can be calculated as 1 minus these two numbers.

The most important issue is usually which phenotypes are associated with the SNP, either directly or indirectly, which can be assessed by summing the probabilities of these two options (or equivalently as 1 minus the probability of the phenotype being unaffected). For example, in the example above, SNP X has a marginal posterior probability of 1-y =z of being associated with the first phenotype.

A quick warning: the prior we use is based on the idea that if a SNP is associated with one of the phenotypes, then it may well be associated with multiple phenotypes. That is, it is relatively permissive of associations with multiple phenotypes, and does not attempt to be "skeptical" of additional associations. In this sense it is more suited to hypothesis generation of additional associations than of hypothesis testing of additional associations. If you want to be more skeptical you will have to change this prior, which will require you to dig into the last output file, below.

```
<test.mph2.mph.txt>

rs      100      200      010      110      210      ...
prior +0.11111 0.00000 +0.11111 +0.05556 +0.02778 ...
rs1   +0.20911 0.00000 -0.08590 +0.08647 -0.11344 ...
rs13  +0.23082 0.00000 -0.10245 +0.10231 -0.11791 ...
```

This last output file contains the log 10 BF computed for each partition. Most users will probably want to ignore this file, but it may be useful if you want to recompute things for the second output file using a different prior. The first row contains the partition identifier. For example, the identifier `210` indicates the partition where the first and the second phenotypes are indirectly (2) and directly (1) affected, respectively, and the third phenotype is unaffected (0). The second row contains the default prior on partitions used by BIMBAM. From the third row, each row lists the SNP name and $\log_{10}$ BF for each partition of the SNP.

References

[1] Bertrand Servin and Matthew Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, 3(7):e114, July 2007.

[2] Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PloS one*, 8(7):e65245, January 2013.

[3] Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nature reviews. Genetics*, 10(10):681–90, October 2009.